Named entity linking

PhD topic - David Kubeša
What is NEL

- Finding names of people, organizations and geographical locations (NER)
- Assigning these entities to knowledge base (EL)
Usage

- Geographical queries
- Time based queries – everybody from 16th century
- Part of the dialogue systems - ability to track entities in conversations
- Cah help machine translation
- Cross language linking - for example from slovak text to czech + slovak kb
- Entity relation
Datasets

- NER defined from 1995
  - Main datasets for German, English, Spanish, Dutch - CoNLL 2002, 2003
  - Czech dataset CNEC - 2007 - UFAL

- NEL
  - First work in 2003
  - Dataset from 2011 - CoNLL + linking AIDA-CONLL

- In our case - Wikidata and Wikipedia - noise
Wikidata

- Structured database of entities corresponding with Wikipedia
- Instances
- Classes
Knowledge base linking challenges

- Entities in the article that are not in the knowledge base
- Same entity that it is not in knowledge base but has multiple occurrences in the text
- Being able to extend knowledge base. Adding by concept and by name
- Ability to create knowledge base from text by clustering and finding relationships
- Ability to link correct pronoun to the entity
- Ability to link correct describing word to the unique entity such as particular mayor
- Priority between different knowledge bases
Adding information about entities to KB

- Looking at the occurrences of the entities in Wikipedia
- Creating descriptions/embeddings
- Parsing Wikipedia
  - Looking for another entities that are not linked – noise problem
  - Possibility of deleting entities that have only one occurrence
  - Another heuristic approaches
- Dealing with missing links in Wikipedia
  - Only first one on page
  - Ambiguity - multiple correct links
  - Consistency of annotation
  - Bigger dataset with noise for training, smaller dataset with as little noise as possible for testing
  - Manual check of the test set
Goals

- Creating reliable dataset for Czech and English from the Wikipedia and Wikidata
- Czech linking system for domestic use
- English system for comparison with SOTA
- Dataset creation algorithm for any combination of languages
- Linking to another language (Czech text to Czech + English knowledge base, etc)
  - Increase of number of named entities
- Cross-lingual transfer during training
  - Useful for languages with limited size of wikipedia
  - Multilingual embeddings
How to

- Processing Wikidata - creating knowledge base
- Component based approach
  - Using UFAL developed NER
  - NN based on SOTA model on top
- End to end system
  - Bert based
  - Multilingual word embeddings
Questions

- Smoke signals preferred, please tell me which location I should look
- Or you can write me